Project Proposal

[ITEC 4305]

Winter 2024

Project Title:	Toronto Fire Incidents: monetary losses prediction using regression models
Supervisor:	Soroush Sheikh Gargar
Student Name:	Ryan Liyanage Derek Anthony Hau Kau Fong Maurílio Tiago Brüning Schmitt
Student ID:	216963654 218387209 220052197
Student Email:	ryanliy@my.yorku.ca derekhkf@my.yorku.ca maurilio@my.yorku.ca

Dataset: https://open.toronto.ca/dataset/fire-incidents/

This dataset includes fire incidents as defined by the Ontario Fire Marshal (OFM) up to December 31, 2022. It contains 43 columns and nearly 30k rows.

1. Objective:

The objective of this project is to estimate dollar loss of fire incidents, using open fire incident data from the city of Toronto. By analysing factors including, but not limited to, the extent of fire, source of fire, ignition material and property use, we aim to provide an accurate estimation of the amount of dollar loss per fire incident. The project seeks to uncover these circumstantial factors that affect the financial impact per every fire incident in Toronto and can offer recommendations to reduce the severity of fire incident impact to fire marshals.

2. Motivation:

Fire incidents are one of the most recurring and devastating problems humans must deal with since ancient times. There is a great dollar loss amount lost due to people displaced, loss of life, and property damage from a fire incident. Understanding the circumstantial events surrounding fire incidents, and how they affect financial impact, can create a new strategy of fire prevention and emergency response allocation. Public infrastructure heavily relies on emergency response and fire fighting to keep citizens safe. Every year, the city of Toronto spends more than half a million dollars on the operating budget for the Toronto Fire Services. [1] This project not only has the potential to provide insight into mitigating economic loss, but also the potential to mitigate human loss of life, and casualties of a fire incident.

3. Related Work:

Lin [2] identified the most significant factors in fire losses using factor analysis, correlation analysis, and regression analysis to analyze 918 cases of residential building fires in Taiwan. The investigation included attributes of occupants and building fire safety, time and spatial attributes of fire occurrence, fire development and egress, fire brigade interventions and the resulting losses. The authors say that the degree of fire severity has the strongest effect on the fire losses, followed by dispatched fire-fighting forces and then fire control time.

Sharma et al. [3] proposed statistical and machine learning frameworks to predict the City of Edmonton fire events. One of the models employed was a Negative Binomial (NB2) regression model to predict Edmonton Fire Rescue Service Events. Fire event and neighborhood features were used from the City of Edmonton's Open Data Portal, and spatial geographical data was used from OpenStreetMap. There was a high fire incident event predictability based on circumstances at the neighborhood level and fire station location level.

4. Methodology

4.1 Data Preprocessing

4.1.1 Data Cleaning

Originally, the dataset had 29,495 rows. Incidents that were identified as false positive *(Final_Incident_Type: 03 - NO LOSS OUTDOOR fire (exc: Sus.arson,vandal,child playing,recycling or dump fires)* will be removed (7,212 rows in total). Furthermore, rows that have *null* values for Estimated Loss (response variable) or Area_of_Origin are going to be deleted too (216 rows), resulting in 21,997 rows.

The dataset has other missing values that may impact our model fitting and results. To handle those missing values, our group will apply the nearest neighbour imputation by using *sklearn.impute.KNNImputer*. The feature of the neighbours will be averaged uniformly or weighted by distance to each neighbour. Furthermore, we will apply the same technique for some features that have also undetermined values such as:

- Status_of_Fire_On_Arrival: 9 - Unclassified
- Extent_Of_Fire: 99 - Undetermined
- Ignition_Source: 999 - Undetermined

Another important task is to identify or remove outliers. By using box plots and scatter plots, it was possible to discover the existence of those data points. A Robust Linear Regression model (Huber Regressor) will be used as an option to deal with the outliers.

4.1.2 Data integration

Integrating data will not be necessary because our model will use only one dataset.

4.1.3 Data reduction

Data reduction will be focused on selecting the best predictors to use in our model. Applying correlation analysis, we will identify the variables which have a strong correlation with the response variable: Kruskal-Wallis Test, Spearman coefficient, Chi-Squared (X^2) Test will be utilized.

A scatterplot matrix between multiple numerical variables in the dataset will also help visualize correlations and patterns.

4.1.4 Data transformation

It will be necessary to encode categorical data numerically by applying one-hot encoding and ordinal encoding.

The response variable (**estimated dollar loss**) will be transformed by using a log transformation.

Because we have features with different scales (e.g., control time in seconds and number of responding apparatus), it will be also necessary to normalize them by using a technique such as Z-score normalization.

4.2 Feature Engineering

Our group will create a new feature (**Control Time**) by calculating *Fire_Under_Control_Time - TFS_Alarm_Time*. This new feature will help the model to understand how long the fire burnt things.

A feature called **Response Time** will be added by calculating the *TFS_Arrival_Time - TFS_Alarm_Time*, to know how long took the first arriving unit to incident.

Another idea is to extract the period (day/night), day of the week, month, year, or season from the timestamps.

4.3 Model Selection

Different models will be used to predict the monetary losses caused by a fire:

4.3.1 Linear Models

1. **Multiple Linear Regression (OLS - Ordinary Least Squares):** it is necessary to assess four conditions (model assumptions): Linearity, Independence, Homoscedasticity, and Normality of Residuals.
2. **Lasso (Least Absolute Shrinkage and Selection Operator):** it is used to improve the prediction accuracy and interpretability of regression models. Applying regularization (L1), we can improve the model generalization.
3. **Elastic-Net:** it is useful when there are correlations among features. It combines L1 and L2 regularization.
4. **Huber Regressor:** it is a powerful tool for robust regression, especially when data contains outliers. It employs L2 regularization.

4.3.2 Ensemble methods

Ensemble methods combine different estimators to improve generalizability over a

single estimator [4].

5. **XGBoost Regressor**: it is an efficient implementation of gradient boosting.

4.3.3 Non-Linear Models

6. **Neural networks (MLP - Multi-layer Perceptron):** MLPRegressor is suitable for regression tasks, where the goal is to predict a continuous numeric value (e.g., predicting house prices, stock prices, or temperature) [4].

4.4 Model Training and Evaluation

Multiple Linear Regression Evaluation: the **general linear F-test** and the associated **null hypothesis** allow us to assess whether the additional parameters in the full model significantly improve the model fit compared to the simpler reduced model.

Mean squared error (MSE) will be used as a metric to measure the accuracy of the models. Adjusted R-squared can be another option.

Residual analysis will help us validate and evaluate our regression model. Analysis of variance (ANOVA) can be applied too.

4.5 Optimization strategies

Hyperparameter tuning, K-Fold Cross-Validation and Regularization techniques will be performed to achieve accurate predictions, ensure robustness and prevent overfitting.

5. Exploratory Data Analysis

At the beginning of our exploratory data analysis, we checked what kind of data types we were dealing with and got two data types: object and float64. We then proceeded to conduct feature engineering; consequently creating Control_Time and Response_Time.

After further analysis, we decided to drop some columns that would not align with our project goals.

When it came to missing data, we found that our dataset contained a lot of null values. Which were dealt with carefully:

1. Firstly, we dropped null rows from Area_of_Origin. The reasoning behind this is because if there is no origin of fire then there is no fire. This was done since we want data that is recorded due to a real fire and not false positives.

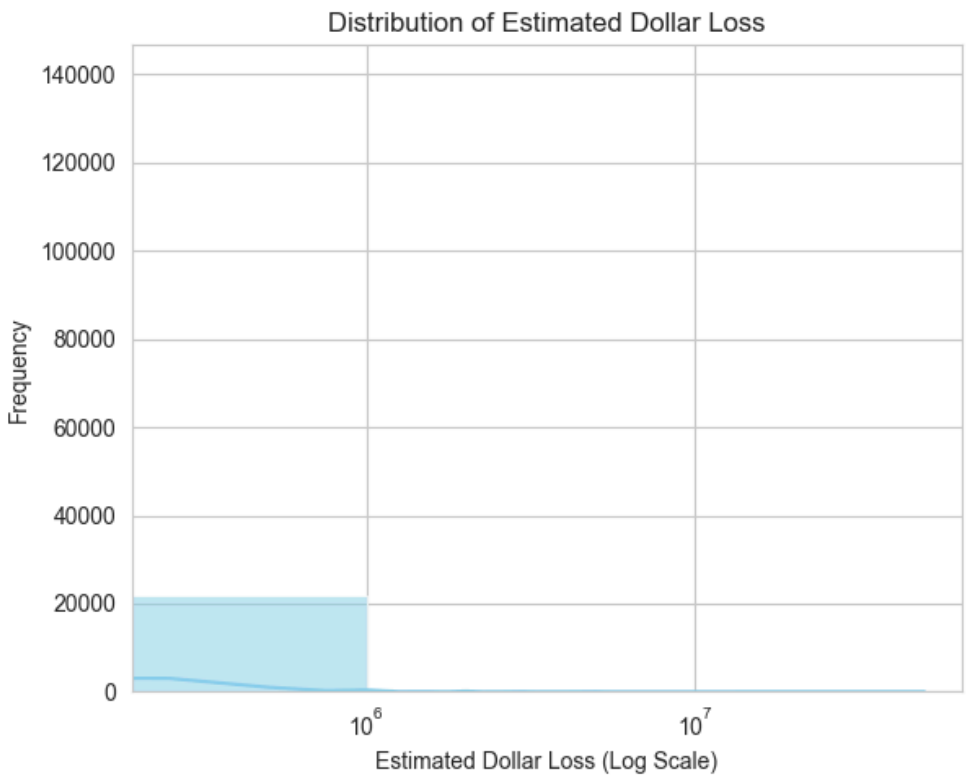
2. We also dropped rows in the column Final_Incident_Type that contained '03 - NO LOSS OUTDOOR' as input. The reason for dropping rows on this condition is because this code meant that no financial loss happened and the other columns were mostly null.

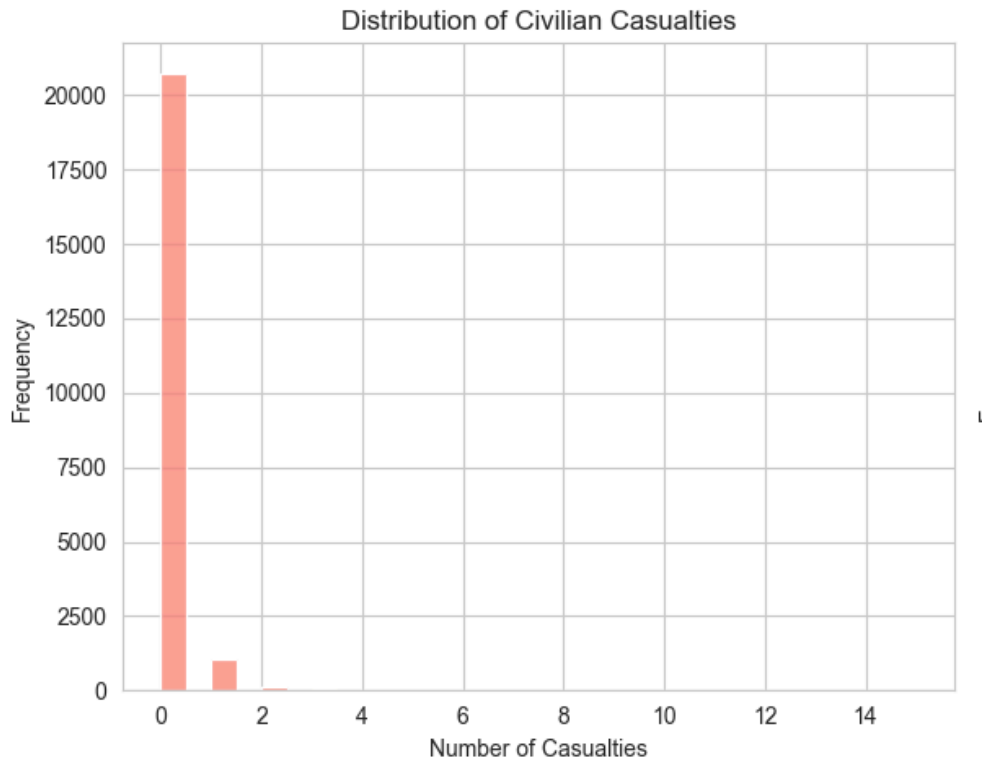
Next, we used a SimpleImputer to impute the remaining rows that still had missing values. Those rows were valuable data but some columns were left empty when the data was recorded.

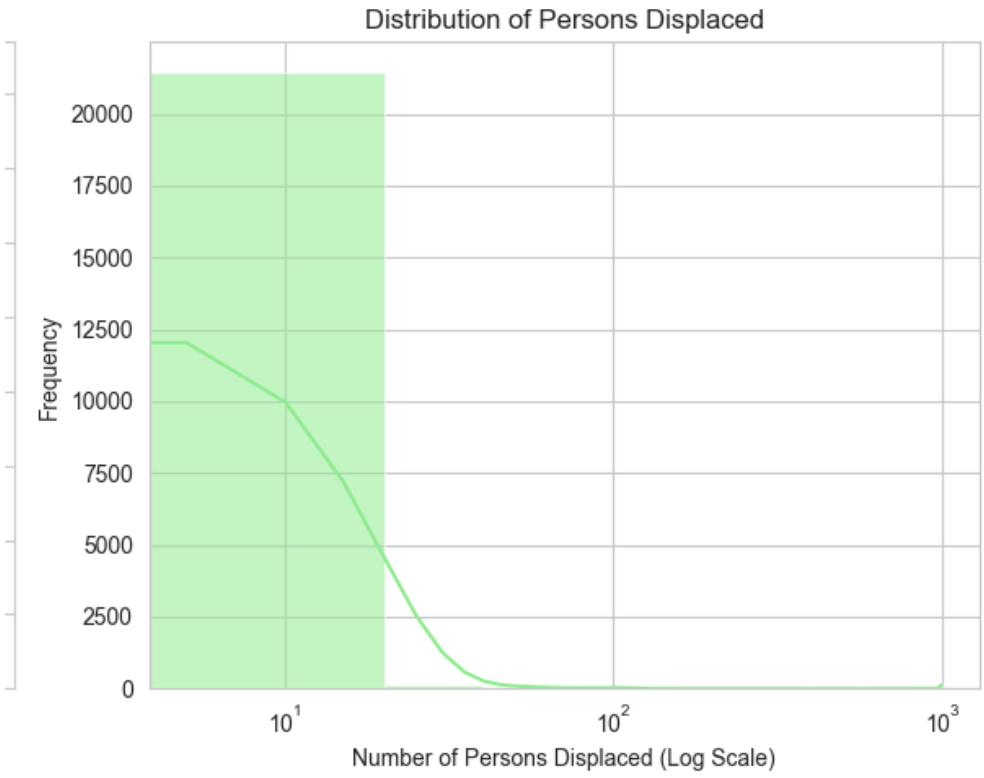
1. For categorical columns, we used a SimpleImputer strategy of most frequent to impute the missing values.

2. For numerical columns, we used a SimpleImputer strategy of median to impute the missing values.

Subsequently, we were able to get our initial visualizations. This resulted in three initial diagrams:


Distribution of Estimated Dollar Loss
140000
120000
100000
80000
60000
40000
20000
0
Frequency
10^6
10^7
Estimated Dollar Loss (Log Scale)

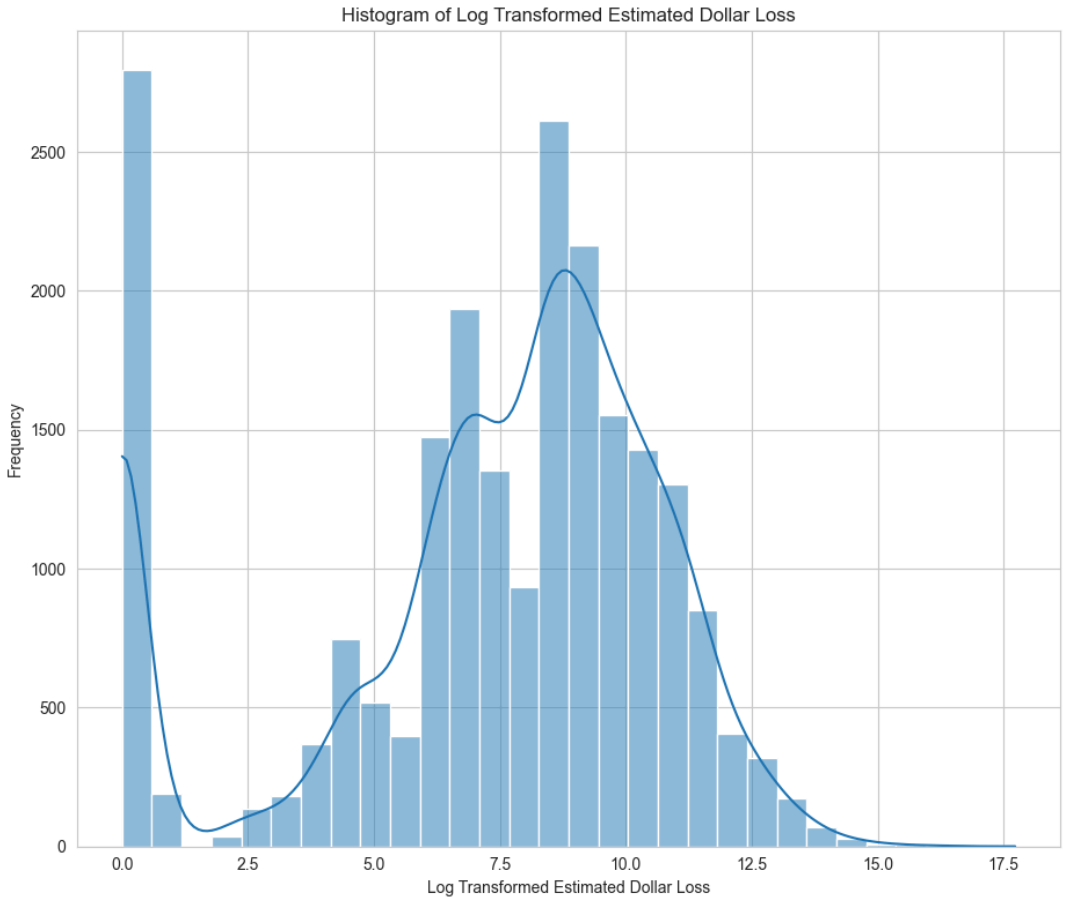

Distribution of Civilian Casualties
20000
17500
15000
12500
10000
7500
5000
2500
0
Frequency
0
2
4
6
8
10
12
14
Number of Casualties



From these, we were able to infer the following:

- Estimated Dollar Loss needs to be readjusted as zero values were not properly dealt with when applying the log scale.
- Number of casualties are in a lot of cases zero, this can be seen from the size of the bar compared to other values.
- Most incidents displace a smaller number of persons, with the frequency decreasing as the number of displaced persons increases.

We then decided to address the issue we had with Estimated_Dollar_Loss by accounting for zero values and shifting the values by 1. After doing so, we plotted a new histogram shown below and we were able to confirm that the log transformation was now successful.

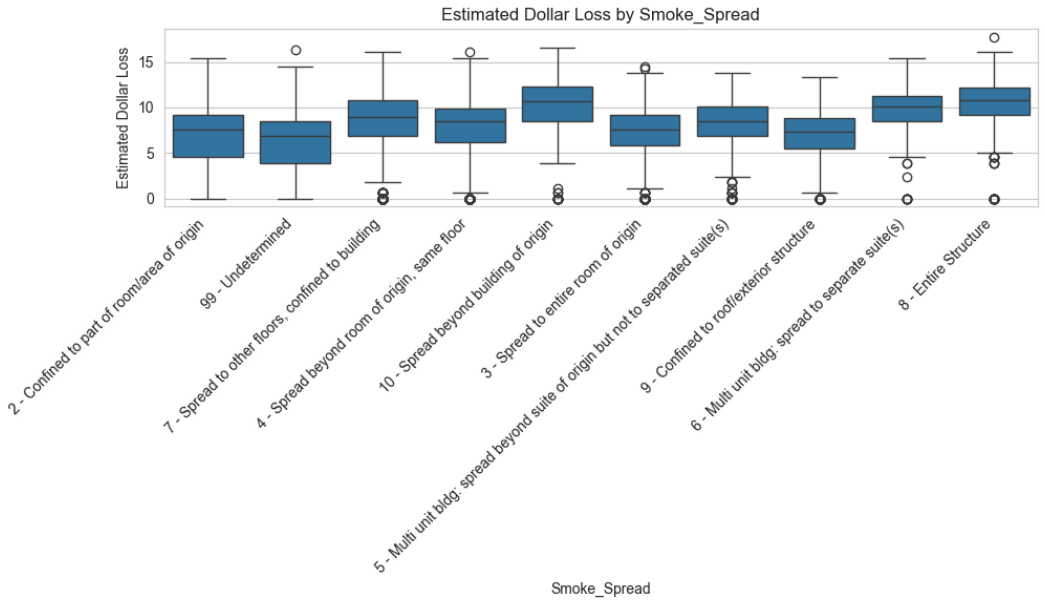


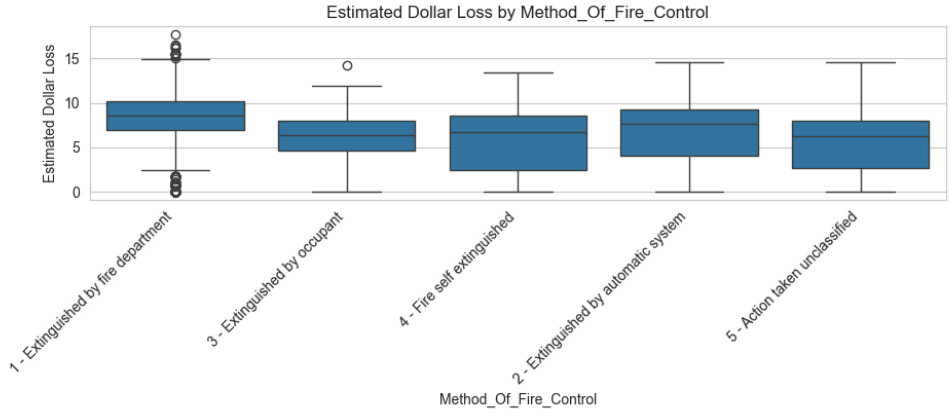
Having a properly log transformed Estimated_Dollar_Loss was crucial to our analysis as it allowed us to better plot further diagrams.

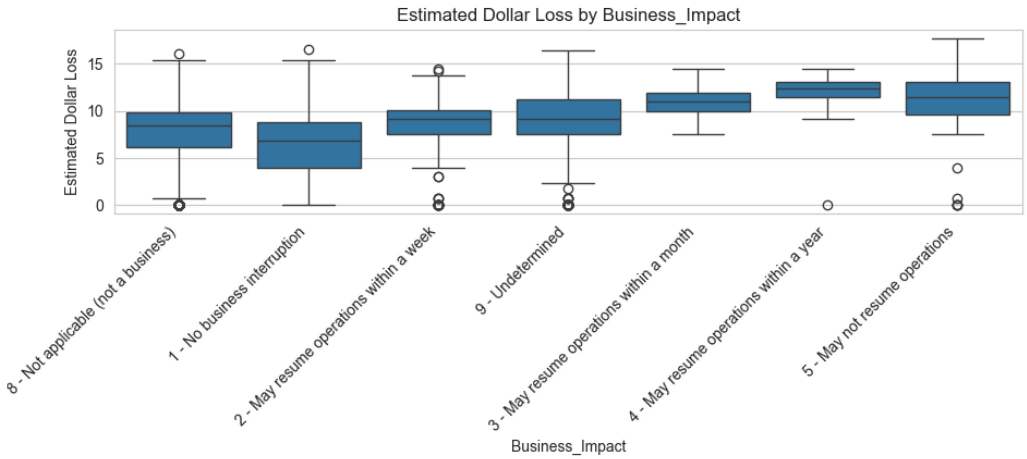
We first visualized diagrams concerning numerical columns.
These diagrams included:

- Box plots
- Correlation Matrix Heatmaps and a bar chart

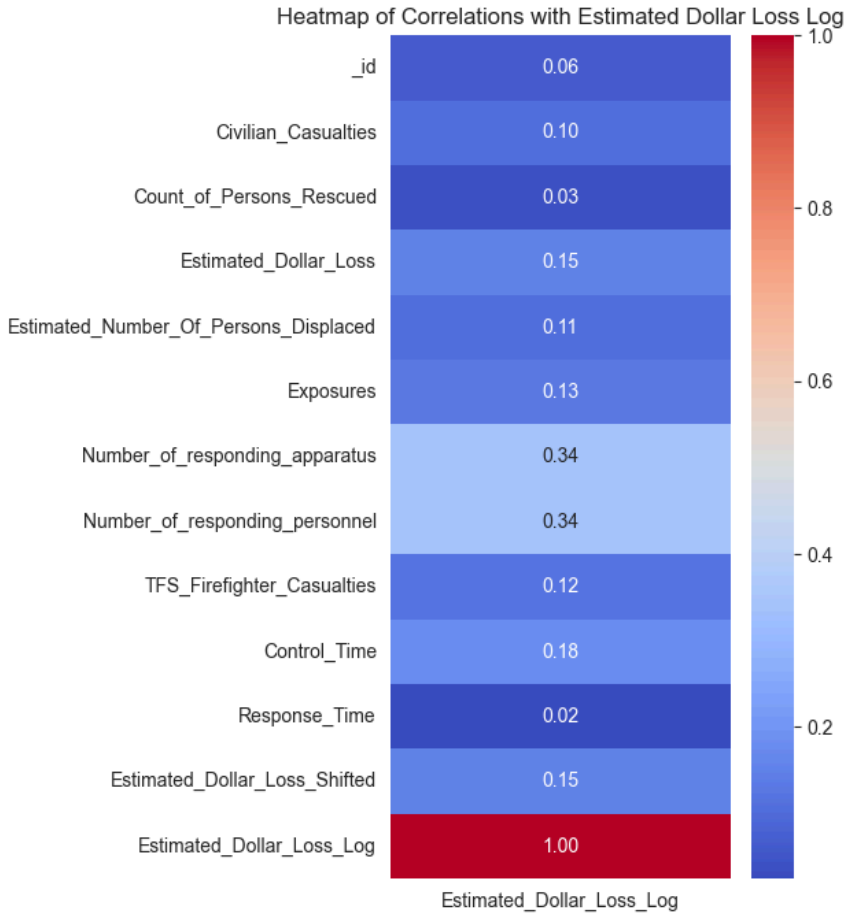
Shown below are a few of the box plots:

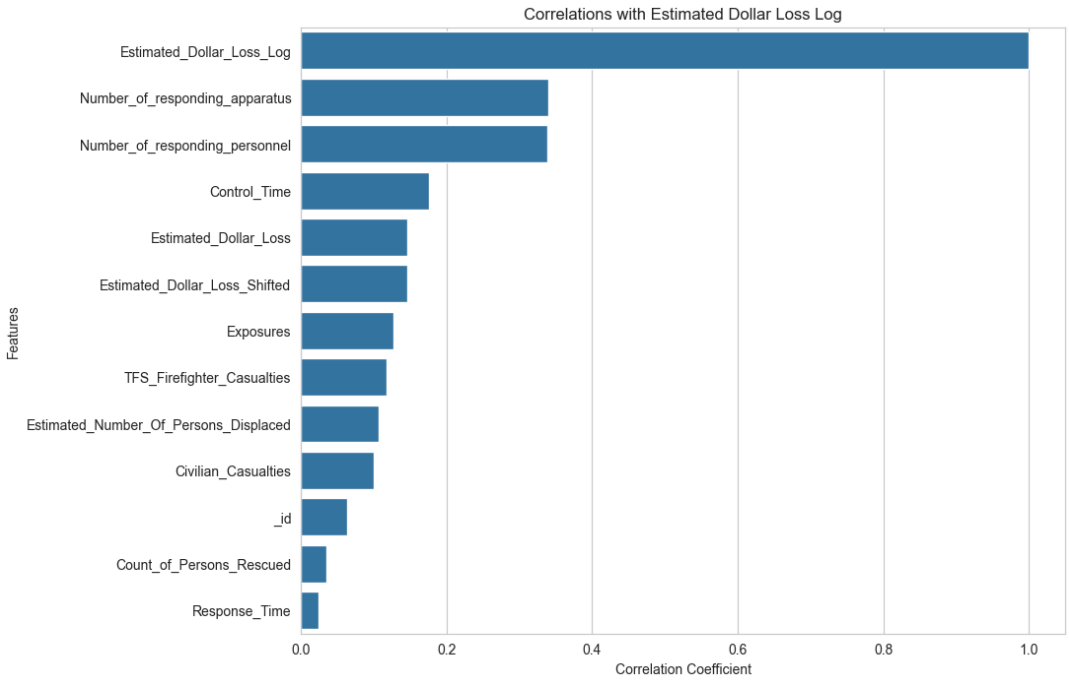

Estimated Dollar Loss by Smoke_Spread
Estimated Dollar Loss
0
5
10
15
2 - Confined to part of room/area of origin
99 - Undetermined
7 - Spread to other floors, confined to building
4 - Spread beyond room of origin, same floor
10 - Spread beyond building of origin
3 - Spread to entire room of origin
5 - Multi unit bldg: spread beyond suite of origin but not to separated suite(s)
9 - Confined to roof/exterior structure
6 - Multi unit bldg: spread to separate suite(s)
8 - Entire Structure
Smoke_Spread


Estimated Dollar Loss by Method_Of_Fire_Control
Estimated Dollar Loss
0
5
10
15
1 - Extinguished by fire department
3 - Extinguished by occupant
4 - Fire self extinguished
2 - Extinguished by automatic system
5 - Action taken unclassified
Method_Of_Fire_Control



And here are the resulting heatmap and bar chart:



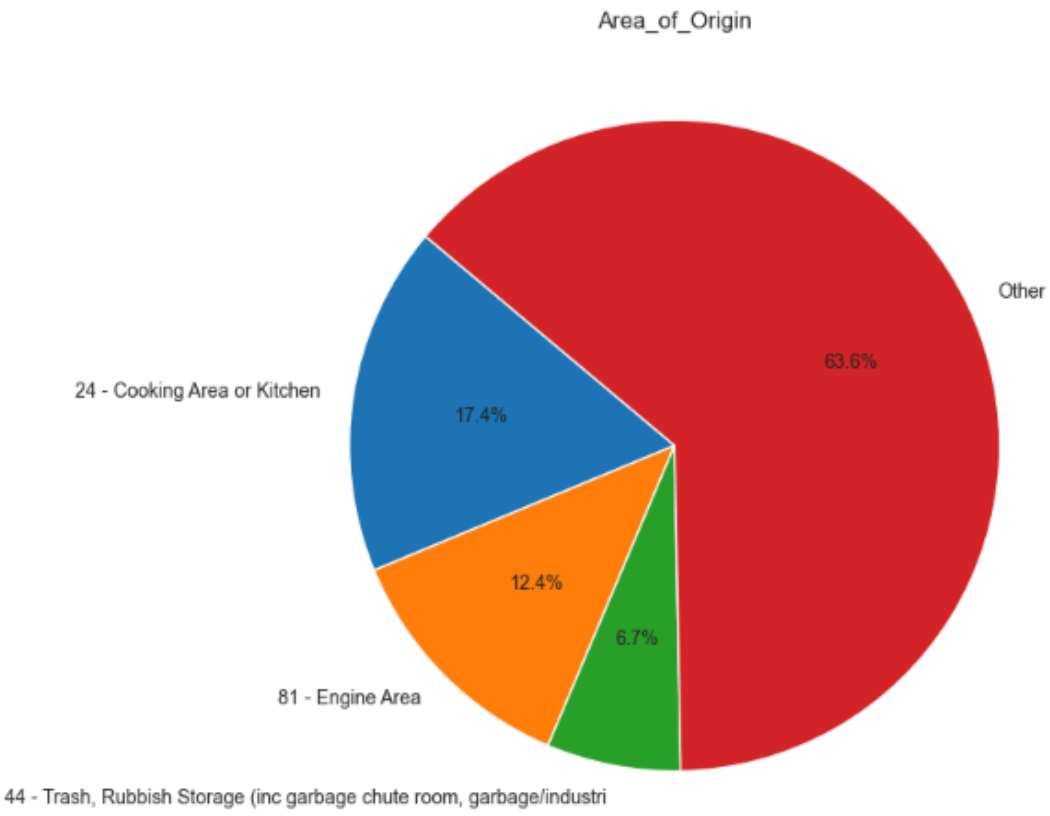


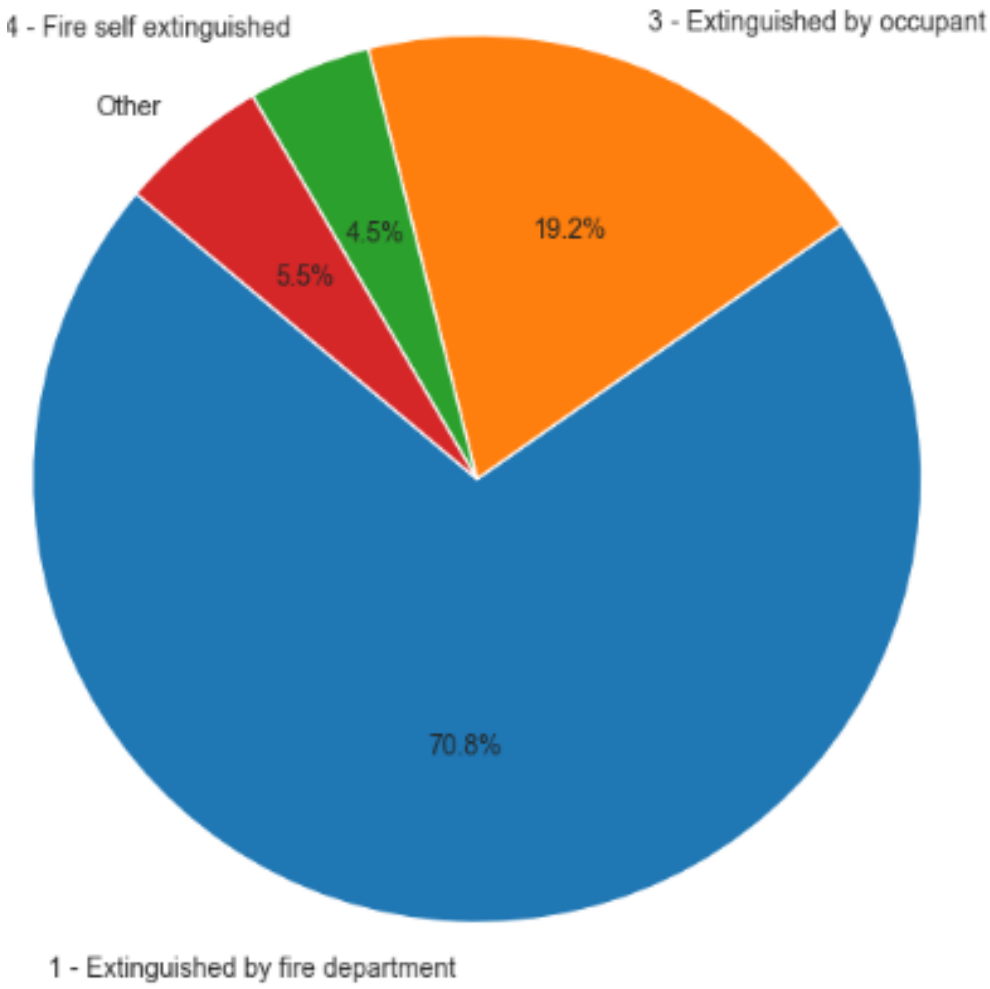
Interpretation from those diagrams:

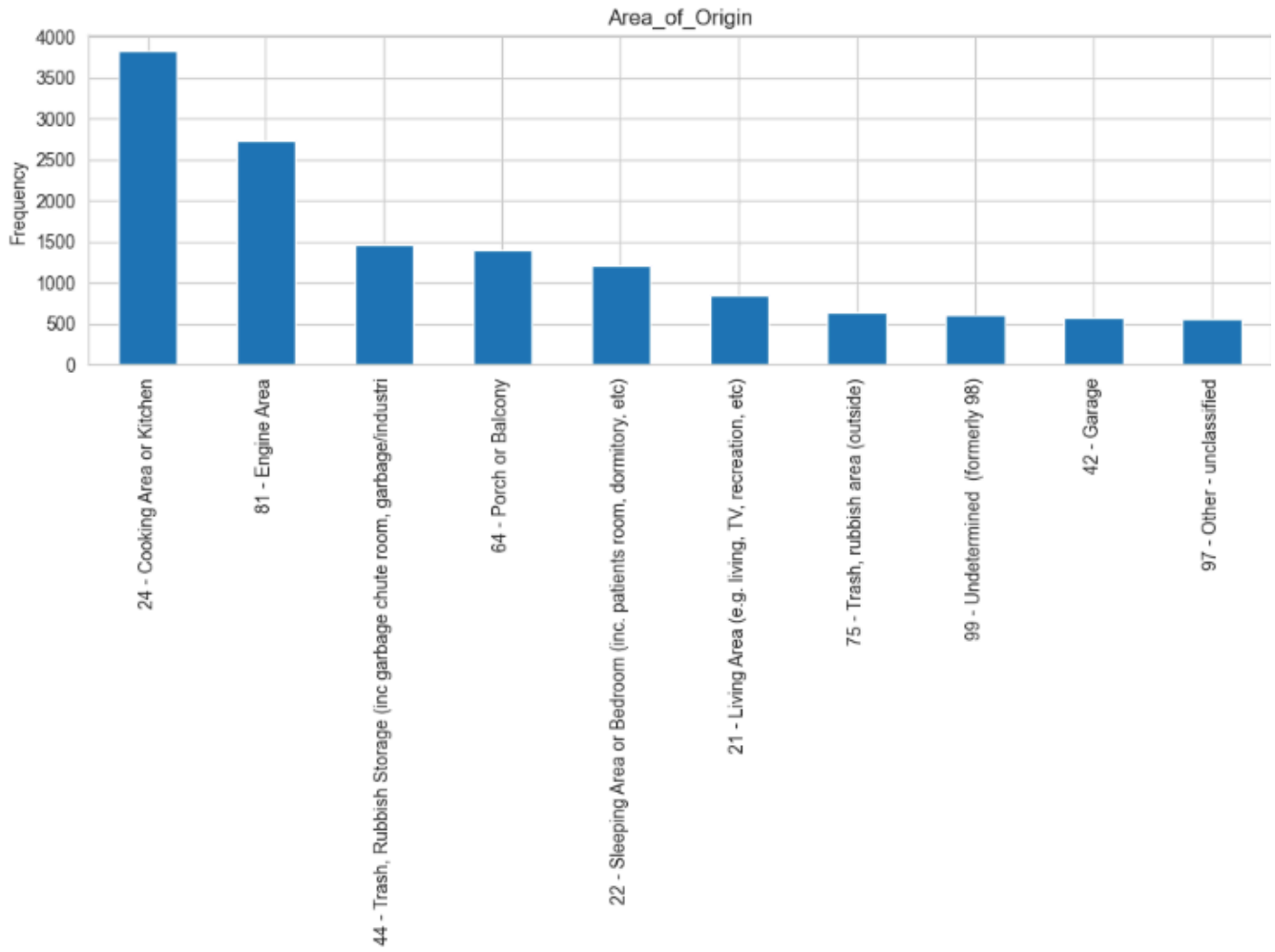
Number_of_responding_apparatus and Number_of_responding_personnel are two promising predictors for now. Because of their correlation coefficient, we might need to take other numerical features to aid in prediction.

We then moved onto categorical columns and we visualized pie charts and bar charts.

Below are some of them:


Area_of_Origin
Other
63.6%
24 - Cooking Area or Kitchen
17.4%
12.4%
6.7%
81 - Engine Area
44 - Trash, Rubbish Storage (inc garbage chute room, garbage/industri


4 - Fire self extinguished
3 - Extinguished by occupant
Other
4.5%
19.2%
5.5%
70.8%
1 - Extinguished by fire department



We finally did some ANOVA tests to prove if categorical features had any effect on Estimated_Dollar_Loss and this is what we found:
Features that had highly significant effects ($p < 0.01$):

- Building_Status
- Business_Impact
- Extent_Of_Fire
- Fire_Alarm_System_Impact_on_Evacuation
- Fire_Alarm_System_Operation
- Fire_Alarm_System_Presence
- Ignition_Source
- Material_First_Ignited
- Method_Of_Fire_Control
- Possible_Cause
- Property_Use
- Smoke_Alarm_at_Fire_Origin

- Related Variables
- Smoke_Spread
- Sprinkler_System_Operation
- Sprinkler_System_Presence

The only feature that had no significant effect was Final_Incident_Type (with a p-value of > 0.05).

This is all the exploratory data analysis that was done so far, further exploration will be done as we progress with the project.

6. Statistical Tests

Source code: https://github.com/RyLiy/TorontoFireIncidents/blob/main/src/preprocessing/hypothesis_testing.ipynb

We can formulate two separate hypotheses to support the validity of the project's objective, that there are circumstantial factors that affect the financial impact of fire incidents.

Hypothesis 1:

- Null Hypothesis (H0): There is no significant association between the number of responding personnel and the estimated dollar loss.
- Alternative Hypothesis (H1): There is a significant association between the number of responding personnel and the estimated dollar loss.

Hypothesis 2:

- Null Hypothesis (H0): There is no significant difference in the estimated dollar loss per fire incident across different materials first ignited.
- Alternative Hypothesis (H1): There is a significant difference in the estimated dollar loss per fire incident across different materials first ignited.

The crux of both hypotheses is whether the presence or absence of specific features (categorized circumstances) can influence the outcome of other features (measured impact)

of fire incidents, particularly the dollar cost lost amount.

Using Spearman's rank-order correlation to validate Hypothesis 1, we were able to obtain a coefficient of 0.56, indicating moderate correlation between responding personnel and estimated dollar loss of each fire incident. With a significance value of 0.01, and a p-value less than 0.001, we can reject the null hypothesis.

Using the Kruskal-Willis Test to validate Hypothesis 2, we were able to compute an H-statistic of 2847.288. With a significance level of 0.01, and the degrees of freedom being one less than the number of different materials that can contribute to fire incidents, we were able to obtain a critical chi-square value of 98.02. As 2847.288 > 98.88, and with a p-value less than 0.01, we can reject the null hypothesis.

Therefore, evidence favouring the alternative hypotheses suggests statistically significant relationships between the circumstances of a fire incident and its impact on estimated dollar loss. As a result, this project will engage in data mining to glean additional insights into the significance of this relationship, to develop predictive models to minimize financial losses by enhancing emergency response and reducing human casualties.

7. Deliverables:

1. **Github repository:** https://github.com/RyLiy/TorontoFireIncidents

    a) The repository shall host Jupyter Notebooks of the preprocessing code, implemented models, pipeline, and model evaluation; raw data, processed data, and documents produced will be saved in respective locations in the repository.

    b) Instructions to build and run the trained models will be outlined in Jupyter Notebooks, in the repository.

2. **Report:** A technical report that will detail exploratory data analysis, data processing techniques, and the systematic machine learning pipelines developed; significance

of results will also be discussed.

3. **Presentation:** A slide deck that will summarize the details and significance of the report produced, along with the motivation of the proposal.

8. Resources:

The project depends on the following resources:

1. **Hardware:**

A device with a multicore CPU no older than 2012, should be able to run Sci-kit learn models and perform statistical calculations, within less than one day. The Python interpreter and Jupyter Notebook kernel can run on nearly all CPU architectures and operating systems.

2. **Software:**

a) An application that can modify and run Jupyter Notebook files.

b) A system with a Python environment installed with dependencies from requirements.txt located in the root of the project Github repository.

3. **Data:**

Datasets shall be procured from Toronto's open data portal, and shall be stored in the Project Github Repository.

9. Impact:

Risk assessment and mitigation of emergencies:

This project can inform the City of Toronto to create a better risk management

framework, by identifying factors that lead to fire incidents, high-risk circumstances can be prevented to avoid severe financial loss. Machine learning models can help predict these factors to allocate resources more effectively.

Optimization of emergency response:

Machine learning models can predict key factors that contribute to higher economic loss. These factors can help first responders prioritize their actions effectively, creating a better fire incident strategy, and better disaster preparedness. Also, based on the Incident Station Areas that resulted in the highest losses, it will be possible to identify the necessity of building new fire stations to decrease response time and control time.

Public safety and well-being:

Ultimately, the findings of this project can contribute to the safety and public well-being of the citizens of Toronto, by reducing the economic loss, and potentially other harmful impacts of fire incidents. By implementing measures based on data-driven decisions, this project can help save lives, protect property, and bolster neighborhood resilience.

10. Milestones:

Milestone	**Target Date**
Proposal document	February 26th
Proposal presentation	February 27th
Model development	March 12th
Model evaluation	March 19th
Final report document	April 1st
Final report slides	April 2nd

Final presentation	April 3rd

11. References:

[1] M. Pegg and C. Williamson, “2024 Budget Notes Toronto Fire Services,” The City of Toronto, Toronto City Hall. Available: https://www.toronto.ca/legdocs/mmis/2024/bu/bgrd/backgroundfile-241810.pdf

[2] Y.-S. Lin, C.-H. Lin, P.-C. Huang
Construction of explanatory fire-loss model for buildings
Fire Safety Journal, 44 (8) (2009), pp. 1046-1052, 10.1016/j.firesaf.2009.07.005

[3] D. P. Sharma et al., “Statistical and Machine Learning Models for Predicting Fire and Other Emergency Events,” arXiv (Cornell University), Feb. 2024, doi: https://doi.org/10.48550/arxiv.2402.09553.

[4] Scikit-learn. (2024). Scikit-learn: Machine Learning in Python. Retrieved from scikit-learn.org